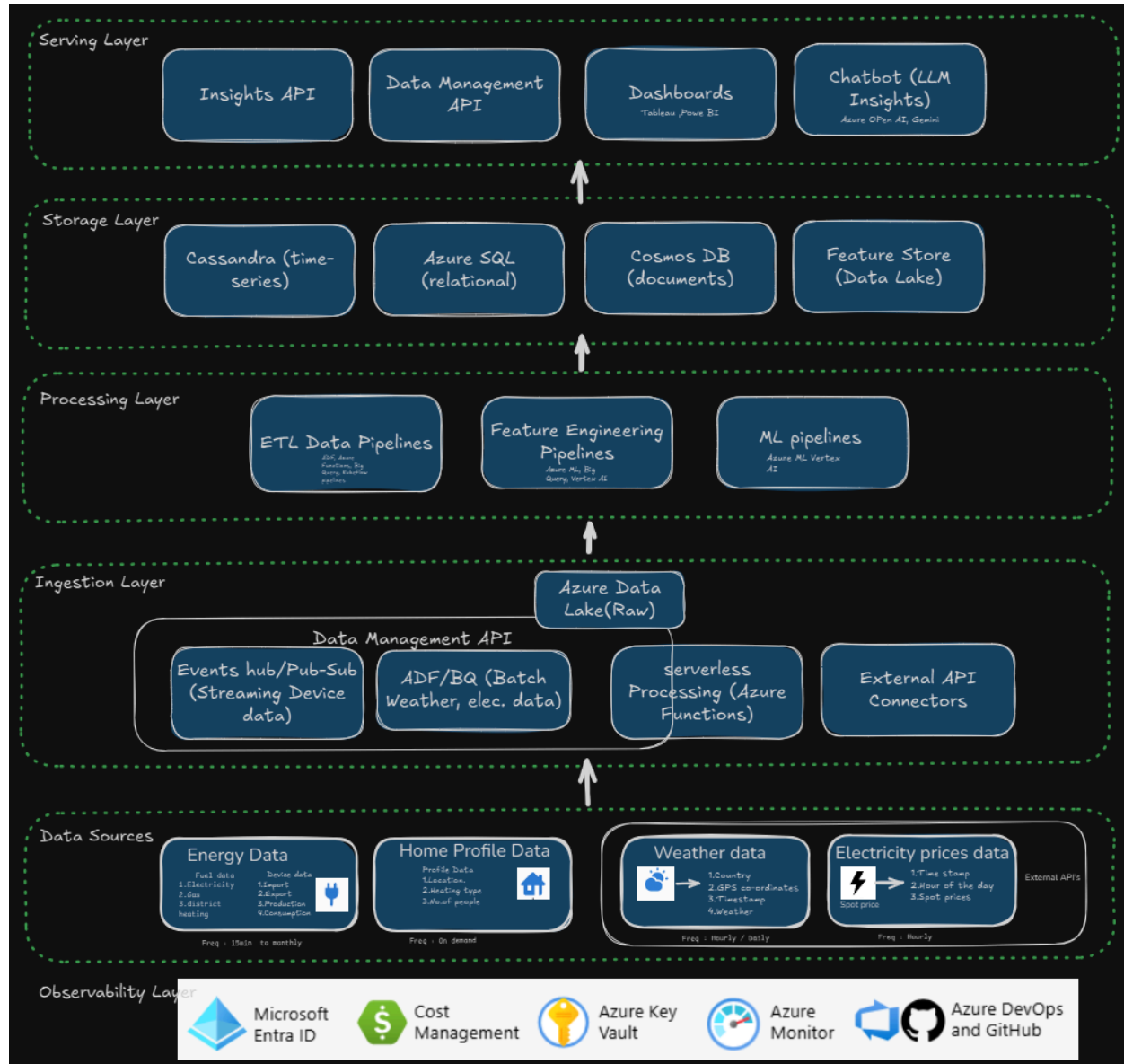# PRD: Eliq Data Platform

1. Objective

Build a **scalable, cloud-native data platform** that:

- Ingests and integrates **energy, home, weather, and market data**.

- Provides a **single source of truth** for analytics and APIs.

- Enables internal teams to **deliver insights quickly and reliably**.

---

2. Key Goals

- **Unified ingestion** from heterogeneous sources.

- **Data quality & governance** at every stage.

- **Faster insights** for internal teams and clients.

- **Modular architecture** that can scale as data grows.

Data Architecture Diagram

Serving Layer
- Insights API
- Data Management API
- Dashboards (Tableau, Powe BI)
- Chatbot (LLM Insights) (Azure OPen AI, Gemini)

Storage Layer
- Cassandra (time-series)
- Azure SQL (relational)
- Cosmos DB (documents)
- Feature Store (Data Lake)

Processing Layer
- ETL Data Pipelines (ADF, dcore KuraCore, Big Query, KubeFlow pipelines)
- Feature Engineering Pipelines (Azure ML, Big Query, Vertex AI)
- ML pipelines (Azure ML Vertex AI)

Ingestion Layer
- Azure Data Lake(Raw)
- Data Management API
  - Events hub/Pub-Sub (Streaming Device data)
  - ADF/BQ (Batch Weather, elec. data)
  - serverless Processing (Azure Functions)
  - External API Connectors

Data Sources
- Energy Data
  - Fuel data
    1. Electricity
    2. Gas
    3. district heating
  - Device data
    1. Import
    2. Export
    3. Production
    4. Consumption
  - Freq : 15min to monthly
- Home Profile Data
  - Profile Data
    1. Location.
    2. Heating type
    3. No.of people
  - Freq : On demand
- Weather data
  1. Country
  2. GPS co-ordinates
  3. Timestamp
  4. Weather
  - Freq : Hourly / Daily
- Electricity prices data
  - Spot price
    1. Time stamp
    2. Hour of the day
    3. Spot prices
  - External API's
  - Freq : Hourly

Observability Layer
- Microsoft Entra ID
- Cost Management
- Azure Key Vault
- Azure Monitor
- Azure DevOps and GitHub

🧩 Component-by-Component Explanation

1. Data Sources

- **Energy Data**: From devices like meters or inverters, capturing import/export/production/consumption. Granularity varies from monthly to 15-min intervals.

- **Home Profile Data**: Stored as structured documents in Cosmos DB, describing physical/occupancy attributes per household.

- **Weather Data**: Hourly data for fixed GPS coordinates via 3rd-party weather API.

- **Electricity Price Data**: Spot prices pulled daily via API for hourly intervals.

- **Internal Insight Services**: Compute on-the-fly results without persisting—used as ephemeral data processors.

---

2. Ingestion Layer

Data Management API handles the data ingestion process from various data sources while External API connectors helps to connect to extract the Weather and elec. data.

- **Azure Data Factory/Big Query**: Handles scheduled pulls for weather and price data.

- **Event Hub/Pub Sub + Azure Functions**: Manages real-time or streaming data from energy devices.

- **Schema validation & transformation**: Light checks and standardization on ingestion.

---

3. Raw Data Lake (Azure Data Lake Gen2/Big Query)

- Central landing zone.

- Immutable, append-only storage.

● Organized by source, date, and location.

---

4. Processing Layer (Azure Functions,Big Query,Vertex AI,Azure ML)

● ETL Pipelines - Cleans, joins, enriches raw data:

  ○ Combines energy with weather and home profiles.

  ○ Aggregates to common time resolutions.

● Feature Engineering Pipelines - Prepares features for modeling.
● Hosts ML models: forecasting, disaggregation, similarity scoring.

---

5. Curated Storage

● **Cassandra**: High-write, distributed time-series store for granular energy data.

● **Azure SQL**: Relational queries for reports and summaries.

● **Cosmos DB**: Retains document-format home profiles.

● **Feature Store**: Centralized, versioned features for model reproducibility.

---

6. Serving Layer

● **Insights API**: Real-time access to computed insights per location.

● **Data Management API**: Secure, structured ingestion endpoint for partners.

● **Tableau/Power BI Dashboards**: Internal tools for Customer Success, Sales, and Ops.
● **Chatbot/LLM based Insights : Enable** users to leverage automated insights using LLMs.

🔁 Integration Approach with Data sources

Source Type	Integration Method	Frequency
Energy data	Event Hub or secure SFTP + Data Factory	Real-time / Daily
Home profile data	Cosmos/Big Query DB (direct sync/API)	On-demand
Weather data	REST API via Azure functions/Data Factory	Hourly / Daily
Market prices	API pull via Functions/Data Factory	Daily
Internal services	Invoked at API request (stateless)	Real-time

🎯 Responsibilities by Team ,Layer,Tools

- **Data Engineering (2-3 people)**: Drives ingestion pipelines, data storage.

- **Data Science (1-2 people)**: Drives ML and feature creation.

- **Platform Engineering(1-2 people)**: Owns APIs and real-time services.

- **Cloud/DevOps(1 people)**: Ensures infrastructure, monitoring, CI/CD.

- **Customer Success(1 people)**: Focused on dashboards and insights delivery.

- **Security/Compliance(1 people)**: Governs access, policies, audits.

RACI Matrix

Key:

- **R** = Responsible (does the work)

- **A** = Accountable (owns the outcome)

- **C** = Consulted (provides input)

- **I** = Informed (kept up to date)

Component / Layer	Data Eng	Data Science	Platform Eng	Cloud / DevOps	Customer Success	Security / Compliance
Data Sources	I	I	I	I	I	I
Event Hub / Streaming	R/A	C	C	C	I	I
ADF / Batch Pipelines	R/A	C	I	C	I	I
Serverless Processing (Azure Functions)	R	C	A	C	I	I
External API Connectors	R	C	A	C	I	I
Data Management API	C	I	R/A	C	I	I
Raw Azure Data Lake	R	C	I	A	I	C
ETL Data Pipelines	R/A	C	I	C	I	I
Feature Engineering Pipelines	C	R/A	I	C	I	I
ML Pipelines	C	R/A	I	C	I	I
Cassandra / Azure SQL / Cosmos DB	R/A	C	I	C	I	C
Feature Store	C	R/A	I	C	I	C
Insights API	C	C	R/A	C	I	I
Dashboards (Power BI / Tableau)	C	C	I	I	R/A	I
Chatbot (LLM Insights)	C	R/A	C	I	I	I
Identity & Access (Microsoft Entra ID)	I	I	I	R/A	I	C
Key Vault / Security Config	I	I	I	R/A	I	C
Cost Management / Azure Monitor	I	I	I	R/A	I	C

🛠️ **How Teams Interact with the Data Platform**

Team	How They Use the Platform	Key Components They Rely On
Data Science	- Consume curated data for analysis and modeling - Build ML pipelines (forecasting, disaggregation, similar homes) - Deploy models exposed via Insights API	Feature Store, Databricks, ML Pipelines, Cosmos DB
Data Engineering	- Build ingestion pipelines (Event Hub, Data Factory) - Maintain ETL workflows - Manage curated storage and schemas	Event Hub, Data Factory, Azure Data Lake, ETL Pipelines, Cassandra, Azure SQL, Cosmos DB
Platform / Backend Eng.	- Develop and manage APIs (Data Management API & Insights API) - Ensure platform uptime and integrations	Data Management API, Insights API, Azure Functions
Customer Success / Analytics	- Use dashboards to view and explain insights to clients - Provide feedback on model accuracy and data quality	Dashboards (Power BI/Tableau), Insights API
Sales / Business Dev.	- Use dashboards and APIs for client demos - Present data -driven insights to prospects	Dashboards, Insights API

Approach if Plan Approved

1. Kick-off sprint with Data Engineering to set up cloud infrastructure.

2. Deliver incremental milestones every 2 weeks (Agile).

3. Conduct reviews after each phase with relevant stakeholders (Engineering, Data Science, Customer Success).

4. Final sign-off once APIs and dashboards meet defined acceptance criteria.

Total Team Size (MVP phase):

- Involving 7–10 people is sufficient for 8–10 weeks of MVP delivery.

Initial User Stories and Tasks

User Story	Tasks	Acceptance Criteria	Review & Sign-off
As a Data Engineer, I can ingest 15-minute energy data into the data lake	- Create Event Hub - Build Functions for validation - Store raw data in Data Lake	- Energy data arrives in `/raw/energy/` partitioned folders - Validation errors logged properly	Reviewed by Data Eng Lead; sign-off by Platform Architect
As a Data Scientist, I can access curated datasets that combine energy, weather & profiles	- Build batch pipelines (ADF/Databricks) - Join energy, weather & home profiles - Create curated tables	- Curated table available in Data Lake and SQL - Data refresh automated daily	Reviewed by Data Science Lead
As a Developer, I can query aggregated energy data via Insights API	- Expose curated data to API - Build endpoint for aggregated consumption queries	- API returns correct aggregation results within SLA	Reviewed by Platform Eng Lead
As a Customer Success Manager, I can view consumption trends on a dashboard	- Connect Power BI to curated datasets - Create basic consumption trends dashboard	- Dashboard shows correct time-series for selected locations	Reviewed by Customer Success Lead

Execution Plan

Phase & Timeline (approx.)	Key Activities	Outcome Deliverables
Week 1–2: Foundation	- Provision Azure Data Lake - Set up Event Hub & Data Factory - Define raw data schemas	Raw landing zone & ingestion setup
Week 3–4: Ingestion	- Build pipelines for energy, weather, market data - Implement validation & logging	Automated data ingestion running
Week 5–6: Processing	- Develop Databricks ETL pipelines - Curate datasets (energy + weather + profiles)	Curated tables ready for consumption
Week 7–8: Serving	- Connect curated data to Insights API-Create dashboards in Power BI	APIs live and dashboards available
Ongoing	- Add feature store - Build ML pipelines - Enhance dashboards	Iterative improvements (V1 → V2)